Names: Nehil Joshi, Harshil Shah, Jashesh Mehta, Hasti Panchal.
Group 36
Subject: CSP 571 Data Preparation & Analysis

Final Project: Data Scientist Salary Analysis

Analyzing Salary Trends and Building Predictive Models

Introduction

- Background:
  In today's competitive job market, salary data for roles such as Data Scientists is invaluable for both job seekers and companies aiming to attract the best talent. This project explores trends and patterns in Data Scientist salaries across different industries, locations, and job titles.
- Objectives:
  The objective of this project is to:
    - Analyze the salary data for Data Scientist roles.
    - Visualize salary trends and correlations between key features.
    - Build predictive models to estimate salaries.
- Dataset Overview:
  The dataset used in this project contains information from 742 job postings for Data Scientist roles. Key features include job titles, company details, skills, and salary ranges.

Data Cleaning and Preprocessing

- Handling Missing Data:
  Missing values in the dataset were handled by replacing invalid values, such as -1, with NaN. The Avg Salary(K) feature was created by averaging the lower and upper salary bounds.
- Exploratory Data Analysis: To understand the dataset better, basic statistics and missing values were explored using df.describe(), df.info(), and df.isnull().sum().

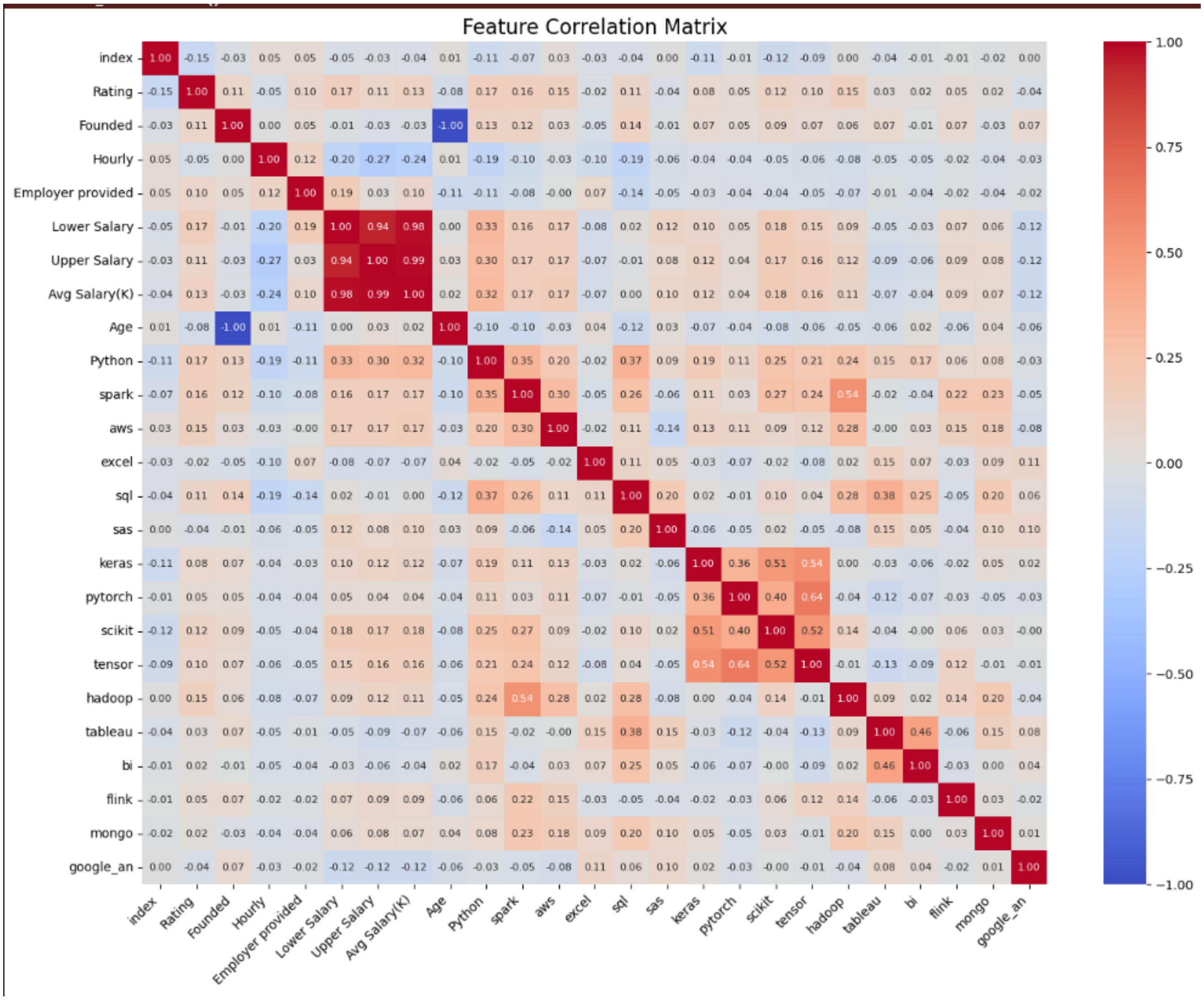
- Code:

```
# Replace -1 values with NaN
df['Rating'].replace(-1, np.nan, inplace=True)
df['Founded'].replace(-1, np.nan, inplace=True)
# Create Avg Salary column
df['Avg Salary(K)'] = (df['Lower Salary'] + df['Upper Salary']) / 2
```

- Correlation Analysis:

  A correlation matrix was calculated to identify relationships between numerical variables, especially between Avg Salary(K) and other features such as Lower Salary and Upper Salary.

  - **Graph: Correlation Heatmap**



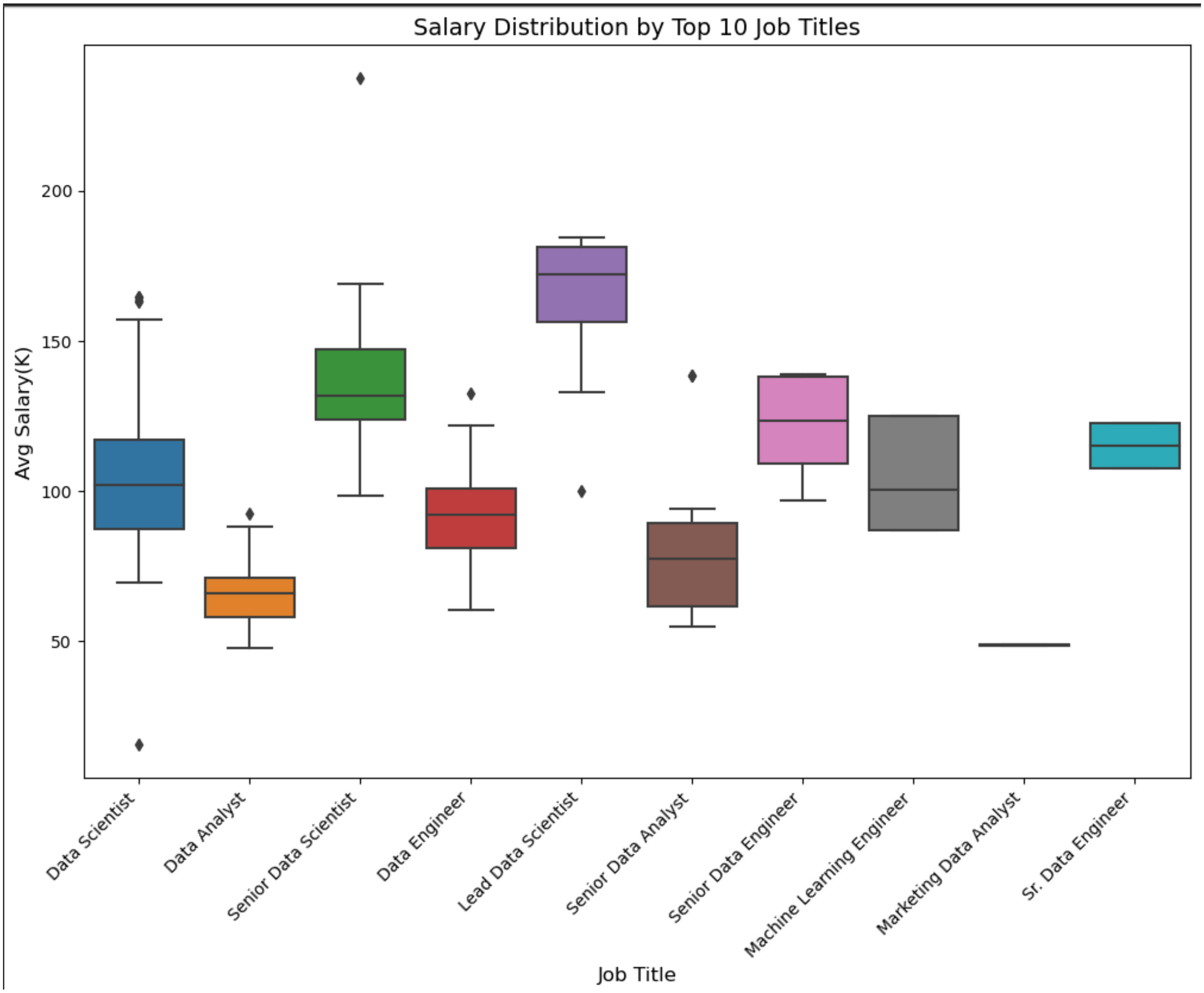
	index	Rating	Founded	Hourly	Employer provided	Lower Salary	Upper Salary	Avg Salary(K)	Age	Python	spark	aws	excel	sql	sas	keras	pytorch	scikit	tensor	hadoop	tableau	bi	flink	mongo	google_an
index	1.00	-0.15	-0.03	0.05	0.05	-0.05	-0.03	-0.04	0.01	-0.11	-0.07	0.03	-0.03	-0.04	0.00	-0.11	-0.01	-0.12	-0.09	0.00	-0.04	-0.01	-0.01	-0.02	0.00
Rating	-0.15	1.00	0.11	-0.05	0.10	0.17	0.11	0.13	-0.08	0.17	0.16	0.15	-0.02	0.11	-0.04	0.08	0.05	0.12	0.10	0.15	0.03	0.02	0.05	0.02	-0.04
Founded	-0.03	0.11	1.00	0.00	0.05	-0.01	-0.03	-0.03	-1.00	0.13	0.12	0.03	-0.05	0.14	-0.01	0.07	0.05	0.09	0.07	0.06	0.07	-0.01	0.07	-0.03	0.07
Hourly	0.05	-0.05	0.00	1.00	0.12	-0.20	-0.27	-0.24	0.01	-0.19	-0.10	-0.03	-0.10	-0.19	-0.06	-0.04	-0.04	-0.05	-0.06	-0.08	-0.05	-0.05	-0.02	-0.04	-0.03
Employer provided	0.05	0.10	0.05	0.12	1.00	0.19	0.03	0.10	-0.11	-0.11	-0.08	-0.00	0.07	-0.14	-0.05	-0.03	-0.04	-0.04	-0.05	-0.07	-0.01	-0.04	-0.02	-0.04	-0.02
Lower Salary	-0.05	0.17	-0.01	-0.20	0.19	1.00	0.94	0.98	0.00	0.33	0.16	0.17	-0.08	0.02	0.12	0.10	0.05	0.18	0.15	0.09	-0.05	-0.03	0.07	0.06	-0.12
Upper Salary	-0.03	0.11	-0.03	-0.27	0.03	0.94	1.00	0.99	0.03	0.30	0.17	0.17	-0.07	-0.01	0.08	0.12	0.04	0.17	0.16	0.12	-0.09	-0.06	0.09	0.08	-0.12
Avg Salary(K)	-0.04	0.13	-0.03	-0.24	0.10	0.98	0.99	1.00	0.02	0.32	0.17	0.17	-0.07	0.00	0.10	0.12	0.04	0.18	0.16	0.11	-0.07	-0.04	0.09	0.07	-0.12
Age	0.01	-0.08	-1.00	0.01	-0.11	0.00	0.03	0.02	1.00	-0.10	-0.10	-0.03	0.04	-0.12	0.03	-0.07	-0.04	-0.08	-0.06	-0.05	-0.06	0.02	-0.06	0.04	-0.06
Python	-0.11	0.17	0.13	-0.19	-0.11	0.33	0.30	0.32	-0.10	1.00	0.35	0.20	-0.02	0.37	0.09	0.19	0.11	0.25	0.21	0.24	0.15	0.17	0.06	0.08	-0.03
spark	-0.07	0.16	0.12	-0.10	-0.08	0.16	0.17	0.17	-0.10	0.35	1.00	0.30	-0.05	0.26	-0.06	0.11	0.03	0.27	0.24	0.54	-0.02	-0.04	0.22	0.23	-0.05
aws	0.03	0.15	0.03	-0.03	-0.00	0.17	0.17	0.17	-0.03	0.20	0.30	1.00	-0.02	0.11	-0.14	0.13	0.11	0.09	0.12	0.28	-0.00	0.03	0.15	0.18	-0.08
excel	-0.03	-0.02	-0.05	-0.10	0.07	-0.08	-0.07	-0.07	0.04	-0.02	-0.05	-0.02	1.00	0.11	0.05	-0.03	-0.07	-0.02	-0.08	0.02	0.15	0.07	-0.03	0.09	0.11
sql	-0.04	0.11	0.14	-0.19	-0.14	0.02	-0.01	0.00	-0.12	0.37	0.26	0.11	0.11	1.00	0.20	0.02	-0.01	0.10	0.04	0.28	0.38	0.25	-0.05	0.20	0.06
sas	0.00	-0.04	-0.01	-0.06	-0.05	0.12	0.08	0.10	0.03	0.09	-0.06	-0.14	0.05	0.20	1.00	-0.06	-0.05	0.02	-0.05	-0.08	0.15	0.05	-0.04	0.10	0.10
keras	-0.11	0.08	0.07	-0.04	-0.03	0.10	0.12	0.12	-0.07	0.19	0.11	0.13	-0.03	0.02	-0.06	1.00	0.36	0.51	0.54	0.00	-0.03	-0.06	-0.02	0.05	0.02
pytorch	-0.01	0.05	0.05	-0.04	-0.04	0.05	0.04	0.04	-0.04	0.11	0.03	0.11	-0.07	-0.01	-0.05	0.36	1.00	0.40	0.64	-0.04	-0.12	-0.07	-0.03	-0.05	-0.03
scikit	-0.12	0.12	0.09	-0.05	-0.04	0.18	0.17	0.18	-0.08	0.25	0.27	0.09	-0.02	0.10	0.02	0.51	0.40	1.00	0.52	0.14	-0.04	-0.00	0.06	0.03	-0.00
tensor	-0.09	0.10	0.07	-0.06	-0.05	0.15	0.16	0.16	-0.06	0.21	0.24	0.12	-0.08	0.04	-0.05	0.54	0.64	0.52	1.00	-0.01	-0.13	-0.09	0.12	-0.01	-0.01
hadoop	0.00	0.15	0.06	-0.08	-0.07	0.09	0.12	0.11	-0.05	0.24	0.54	0.28	0.02	0.28	-0.08	0.00	-0.04	0.14	-0.01	1.00	0.09	0.02	0.14	0.20	-0.04
tableau	-0.04	0.03	0.07	-0.05	-0.01	-0.05	-0.09	-0.07	-0.06	0.15	-0.02	-0.00	0.15	0.38	0.15	-0.03	-0.12	-0.04	-0.13	0.09	1.00	0.46	-0.06	0.15	0.08
bi	-0.01	0.02	-0.01	-0.05	-0.04	-0.03	-0.06	-0.04	0.02	0.17	-0.04	0.03	0.07	0.25	0.05	-0.06	-0.07	-0.00	-0.09	0.02	0.46	1.00	-0.03	0.00	0.04
flink	-0.01	0.05	0.07	-0.02	-0.02	0.07	0.09	0.09	-0.06	0.06	0.22	0.15	-0.03	-0.05	-0.04	-0.02	-0.03	0.06	0.12	0.14	-0.06	-0.03	1.00	0.03	-0.02
mongo	-0.02	0.02	-0.03	-0.04	-0.04	0.06	0.08	0.07	0.04	0.08	0.23	0.18	0.09	0.20	0.10	0.05	-0.05	0.03	-0.01	0.20	0.15	0.00	0.03	1.00	0.01
google_an	0.00	-0.04	0.07	-0.03	-0.02	-0.12	-0.12	-0.12	-0.06	-0.03	-0.05	-0.08	0.11	0.06	0.10	0.02	-0.03	-0.00	-0.01	-0.04	0.08	0.04	-0.02	0.01	1.00

Exploratory Data Analysis (EDA)

- Salary Distribution by Job Titles:

  The distribution of salaries by job title was explored using a box plot to highlight the variation in salaries for different roles. Job titles like Senior Data Scientist and Machine Learning Engineer exhibited higher median salaries compared to roles like Data Analyst.

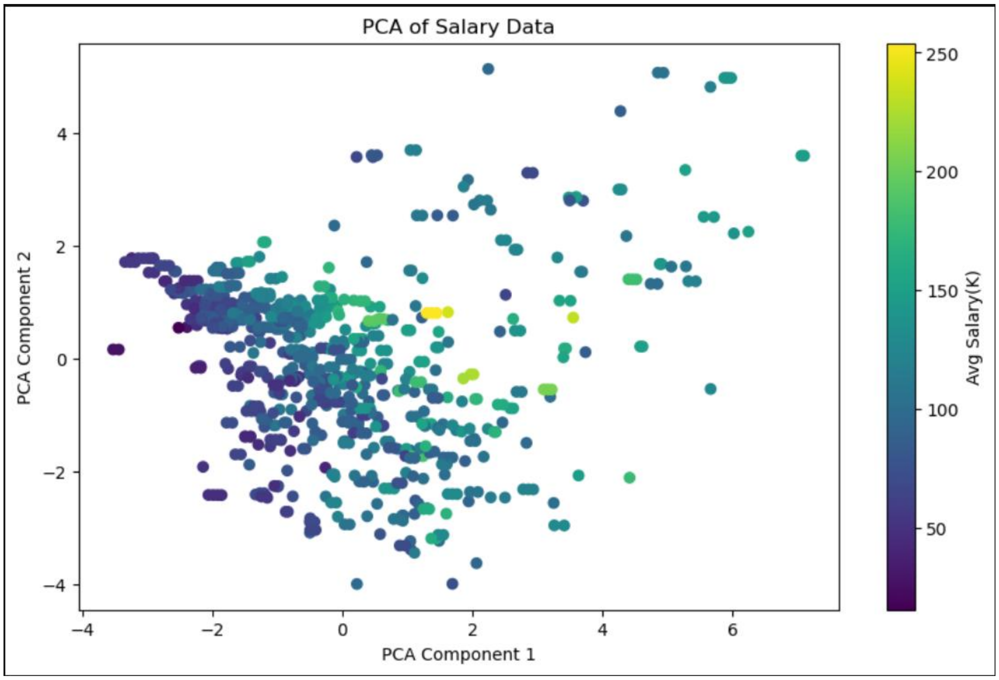
  - **Graph: Salary Distribution by Top 10 Job Titles**



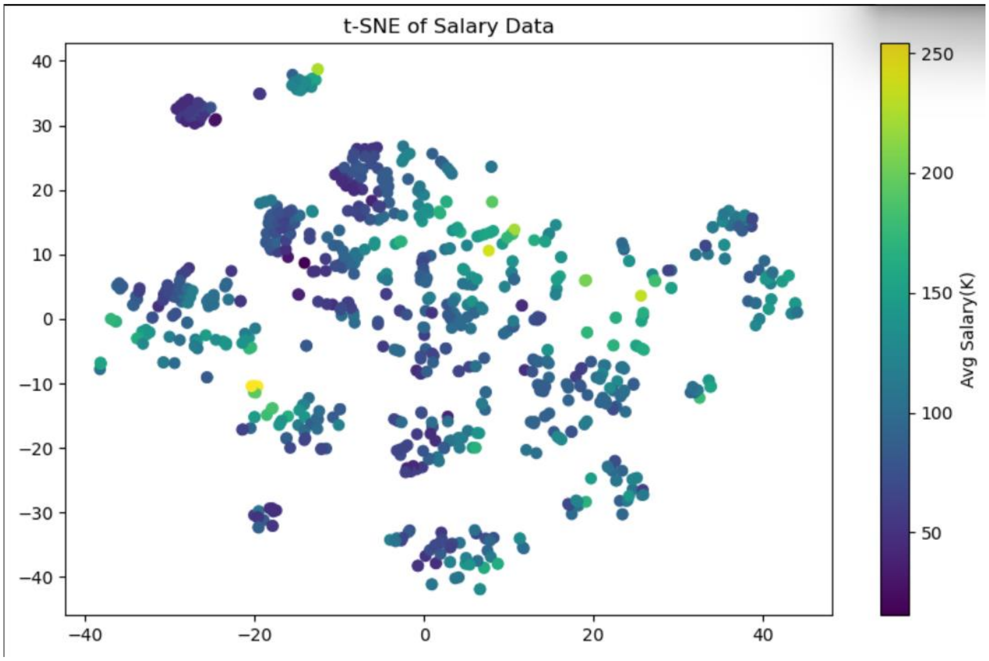
- Dimensionality Reduction:

  PCA and t-SNE were applied to reduce the high-dimensional feature space into two dimensions to visualize the data's underlying structure.

  - **Graph: PCA of Salary Data**

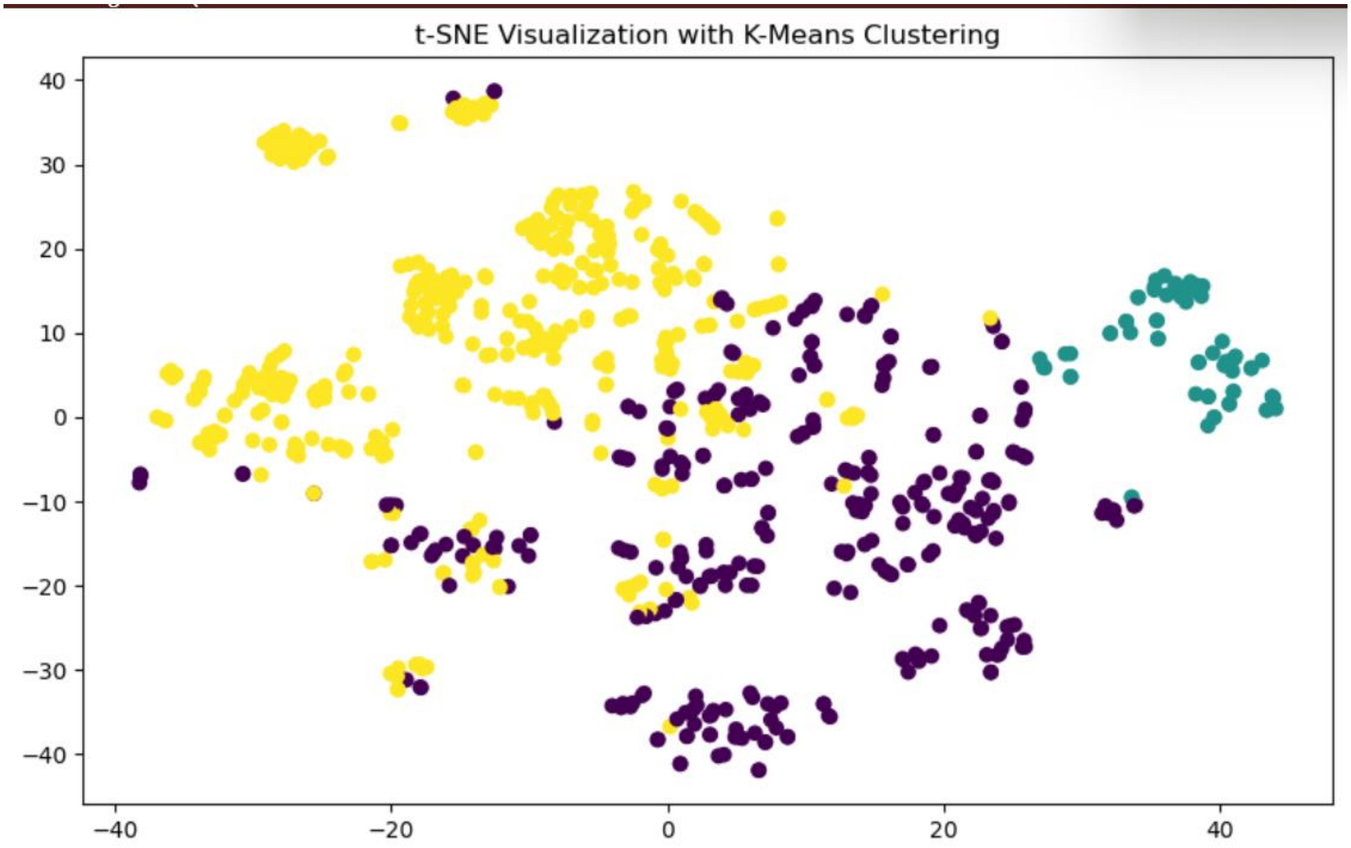


  - **Graph: t-SNE of Salary Data**



Clustering and Predictive Modeling

- Clustering:
  K-means clustering was applied to segment the dataset into three clusters based on the features. The t-SNE visualization with clusters helps understand how data points are grouped.
  - **Graph: t-SNE with K-Means Clusters**



- Predictive Modeling:
  Several machine learning models were trained to predict salaries, including:
  - **Linear Regression**
  - **Lasso and Ridge Regression**
  - **Gradient Boosting Regressor**

After training, the models were evaluated using metrics like R^2, Mean Absolute Error (MAE), and Mean Squared Error (MSE).

  - **Code for Model Training (Example: Linear Regression):**

    ```
    model = LinearRegression()
    model.fit(X_train_selected, y_train)
    y_pred = model.predict(X_test_selected)
    print("MAE:", mean_absolute_error(y_test, y_pred))
    print("MSE:", mean_squared_error(y_test, y_pred))
    print("R2 Score:", r2_score(y_test, y_pred))
    ```

- Model Results:
  - Linear Regression achieved an R^2 of 0.82.
  - Lasso and Ridge regularization models were used to avoid overfitting, with R^2 scores of 0.83 and 0.81, respectively.
  - Gradient Boosting outperformed all, with an R^2 of 0.88, making it the best model for salary prediction.

Insights and Conclusions

- Key Insights:
  - Specialized roles such as Senior Data Scientist and Machine Learning Engineer tend to offer higher salaries.
  - Data from large companies, particularly in tech and finance, indicates higher salary ranges.
  - Skills such as Python, SQL, and cloud technologies like AWS are highly correlated with higher salaries.
- Recommendations for Job Seekers:
  - Job seekers should focus on developing in-demand technical skills.
  - Targeting specialized roles like Senior Data Scientist or Machine Learning Engineer can significantly increase earning potential.
- Conclusion: This project successfully explored and analyzed Data Scientist salaries. By visualizing key trends and building predictive models, it provides insights that can assist both job seekers and employers in understanding salary dynamics. The Gradient Boosting model proved to be the most effective in predicting salaries, offering a robust tool for future salary estimations.